Universität Augsburg

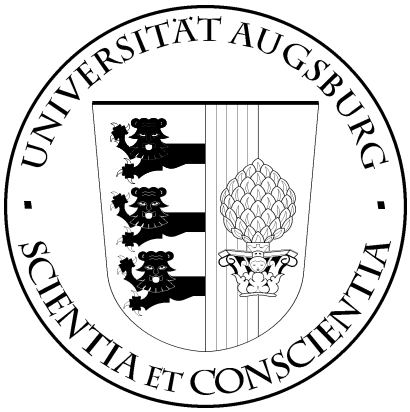
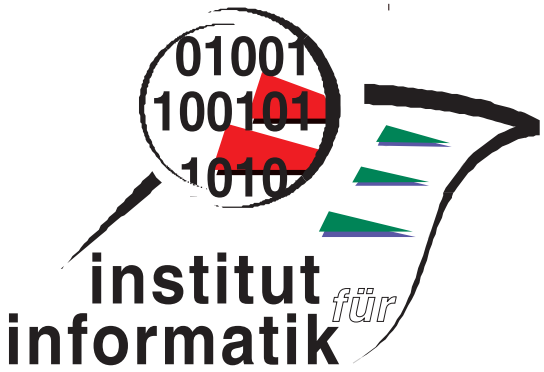


Image Retrieval on Large-Scale Image Databases

E. Hörster, R. Lienhart, M. Slaney

Report 2007-05 April 2007



Institut für Informatik
D-86135 Augsburg

Image Retrieval on Large-Scale Image Databases

Eva Hörster
Multimedia Computing Lab
University of Augsburg
Augsburg, Germany
hoerster@informatik.uni-augsburg.de

Rainer Lienhart
Multimedia Computing Lab
University of Augsburg
Augsburg, Germany
lienhart@informatik.uni-augsburg.de

Malcolm Slaney
Yahoo! Research
Santa Clara, CA 95054
USA
malcolm@ieee.org

ABSTRACT

Online image repositories such as Flickr contain hundreds of millions of images and are growing quickly. Along with that the needs for supporting indexing, searching and browsing is becoming more and more pressing. In this work we will employ the image content as a source of information to retrieve images. We study the representation of images by Latent Dirichlet Allocation (LDA) models for content-based image retrieval. Image representations are learned in an unsupervised fashion, and each image is modeled as the mixture of topics/object parts depicted in the image. This allows us to put images into subspaces for higher-level reasoning which in turn can be used to find similar images. Different similarity measures based on the described image representation are studied. The presented approach is evaluated on a real world image database consisting of more than 246,000 images and compared to image models based on probabilistic Latent Semantic Analysis (pLSA). Results show the suitability of the approach for large-scale databases. Finally we incorporate active learning with user relevance feedback in our framework, which further boosts the retrieval performance.

1. INTRODUCTION

Nowadays there exist online image repositories containing hundreds of millions of images of all kinds of quality, size and content. One example of such an image repository is Flickr™. These repositories grow day by day making techniques for navigating, indexing, and searching prudent. Currently indexing is mainly based on manually entered tags and/or individual and group usage patterns. Manually entered tags, however, are very subjective and not necessarily referring to the shown image content. A good example, for instance is the tag "Christmas" in Flickr. Only a very small fraction of the images depict the religious event as one might expect. Instead the tag often denotes the time and date of creation. Thus thousands of vacation and party photos pop up with no real common theme. This subjectivity and ambiguity of tags makes image retrieval based on manually entered tags difficult.

In this work we employ a different source of information to retrieve images: the image content. Recently advanced generative models originally developed for statistical text modeling in large document collections such as probabilistic Latent Semantic Analysis (pLSA) [7] and Latent Dirichlet Allocation (LDA) [4] have been introduced and repurposed for image content analysis tasks such as scene classification [8] and object recognition [13]. Documents are modeled as mixtures of intermediate (hidden) topics (also called aspects) under the assumption of a bag-of-words document representation. Applied to visual tasks, the mixture of hidden topics refers to the degree to which a certain object/scene type is contained in the image. In the ideal case, this gives rise to a low-dimensional description of the coarse image content and thus enables retrieval in very large databases.

Given unlabeled training images, the probability distributions of the above mentioned models are estimated in a completely unsupervised fashion. The pLSA model has been shown to work in image similarity search tasks in large real world databases [9]. The LDA model is closely related to the pLSA model, but provides a completely generative model and therefore overcomes some problems of the pLSA. Thus the suitability of LDA models to solve the image retrieval problem is studied in this paper. Our evaluation is based on a real world database consisting of more than 246,000 images downloaded from Flickr. The resulting image database was not cleaned nor preprocessed in any way to increase consistency. Retrieval results are evaluated purely based on image similarity as perceived by ordinary users.

By definition query-by-example methods are only able to find images of similar content independent of the precise query concept a user has in his mind. Retrieval results are improved by user relevance feedback since the feedback refines the precise query concept of the user. Thus we combine one of the best active learning approaches [15] with the LDA image representation and a novel preprocessing scheme for data selection in order to improve retrieval results. Performance is evaluated again based on user studies.

1.1 Related Work

Recently a few research groups have started to use probabilistic text models [4, 7] for visual retrieval tasks. In this new approach, each image is modeled as a mixture of hidden topics, which in turn model the co-occurrence of so called visual words inside and across images. These models have been successfully applied and extended to scene classification [5, 8, 12] and object categorization [6, 13, 16]. Varia-

tions of latent space models have also been applied to the problem of modeling annotated images [2, 3].
In the visual domain, so far these aspect models are mostly applied to relatively small, carefully selected image databases ranging from a few hundred to a few thousand images. Those databases are far from being representative for realistic retrieval tasks on large-scale databases. Our previous work [9] shows that the use of pLSA models (i.e., the topic distribution of the images) improves retrieval performance on large-scale real world image database. The work centered on finding 'suitable' visual words and we will build on these insights when computing the visual vocabulary.
In this work we will combine the generative LDA model [4] with a large-scale real world image retrieval task. The work was inspired by previous work [17] that uses LDA models to improve information retrieval.

1.2 Contributions

The main contributions of this paper are:

- We explore the application of LDA models for content based image retrieval and judge its suitability by user studies on a real world, large-scale database with more than 246,000 images.
- We evaluate various parameter settings and different distance measures for similarity judgment. In addition, we perform a competitive comparison with the pLSA-based image representation.
- We apply an active learning algorithm [15] to the LDA-based image representation. Retrieval results are further improved by means of a novel data selection method that prunes the set of candidate images used during active learning.

The paper is organized as follows. Section 2 describes the LDA-based image representation. We outline the visual word computation and review the LDA model. Then we present different similarity measures for example-based retrieval based on the LDA representation. Experimental results of the proposed retrieval system on the complete image database are shown in Section 3. Section 4 outlines the combination of LDA-based image features and an active learning algorithm. Modifications with respect to the original algorithm are described and experimentally evaluated. Section 5 concludes the paper.

2. LDA-BASED IMAGE RETRIEVAL

2.1 Image Representation

Latent Dirichlet Allocation (LDA) [4] is a generative probabilistic model developed for collections of text documents. It represents documents by a finite mixture over latent topics, also called hidden aspects. Each topic in turn is characterized by a distribution over words. In this work our aim is to model image databases not text databases, thus our documents are images and topics correspond to objects depicted in the images. Most importantly LDA allows us to represent an image as a mixture of topics, i.e. as a mixture of multiple objects.
The starting point for building an LDA model is to first represent the entire corpus of documents by a term-document co-occurrence table of size $M \times N$. M indicates the number of documents in the corpus and N the number of different words occurring across the corpus. Each matrix entry stores the number of times a specific word (column index) is observed in a given document (row index). Such a representation ignores the order of words/terms in a document, and is commonly called a bag-of-words model.
When applying those models to images, a finite number of elementary visual parts, called visual words, are defined in order to enable the construction of the co-occurrence table. Then each database image is searched for the occurrences of these visual words. The word occurrences are counted, resulting in a term-frequency vector for each image document. The set of term-frequency vectors constitutes the co-occurrence table of the image database. Since the order of terms in a document is ignored, any geometric relationship between the occurrences of different visual words in images is disregarded.
A finite number of hidden topics is then used in the LDA to model the co-occurrence of (visual) words inside and across documents/images. Each occurrence of a word in a specific document is associated with one unobservable topic. Probability distributions of the visual words given a hidden topic as well as probability distributions of hidden topics given the documents are learned in a complete unsupervised manner.

2.1.1 Visual Words Computation

The first step in computing the observable co-occurrence matrix of words in images is to compute a visual vocabulary consisting of N visual words. This is usually derived by vector quantizing automatically extracted local image descriptors. In this work the well-known SIFT features [10] are chosen as local image descriptors. They are computed in two steps: A sparse set of interest points is detected at extremas in the difference of Gaussian pyramid and a scale, position and orientation are assigned to each interest point. Then we compute a 128-dimensional gradient-based feature vector from the local grayscale gradient neighborhood of each interest point in a scale and orientation invariant manner.
Most works perform k-means clustering on local image features and keep the means of each cluster as visual words. In our previous work [9] we investigated three different techniques for computing visual words from local image features for large-scale image databases such as Flickr. Surprisingly, clustering based on subsets of features derived from images with the same tags did not improve performance. This may be the result of inconsistent labeling as we can often see in community databases such as Flickr. We use the best performing technique [9] – merging the results of multiple k-means clustering on non-overlapping feature subsets – in this work for visual word computation. Therefore relatively small sets (compared to the entire number of features in all 246,000 images) of features are selected randomly from all features. Then k-means clustering is applied to each subset and the derived visual words of each subset are amalgamated into the vocabulary. This approach is more efficient with respect to runtime than determining all clusters from one large set of features.
Given the vocabulary, we represent each image I_d as consisting of N_d visual words by replacing each detected feature vector by its most similar visual word, defined as the closest word in the 128-dimensional vector space.

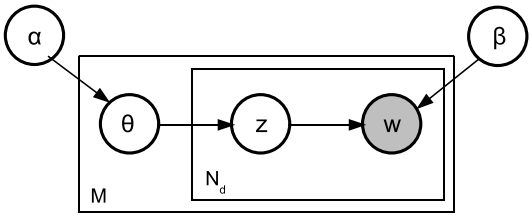


Figure 1: Graphical representation of LDA model (M denotes the number of images in the database and N_d the number of visual words in image I_d). The shadowed node denotes the observable random variable w for the occurence of a visual word, z denotes the topic variable and θ the topic mixture variable.

2.1.2 LDA Model

Each image I_d is represented as a sequence of N_d visual words w_n, and written $\mathbf{w_d} = \{w_1, w_2, ..., w_{N_d}\}$. In an LDA model [4], the process of generating such an image is described as follows:

- Choose a K-dimensional Dirichlet random variable $\theta \sim Dir(\alpha)$, where K denotes the finite number of topics in the corpus.
- For each of the N_d words w_n:
  - Choose a topic $z_n \sim Multinomial(\theta)$
  - Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n

The grahical representation of the LDA model is shown in Figure 1. M indicates the number of images in the entire database and N_d denotes the number of visual words in image I_d.
The likelihood of an image I_d according to this model is given by:

$$p(\mathbf{w_d}|\alpha, \beta) = \int p(\theta|\alpha) \prod_{n=1}^{N_d} (\sum_{j=1}^{K} p(z_j|\theta) p(w_n|z_j, \beta)) d\theta \quad (1)$$

The probability of a corpus/database is the product of the marginal probabilities of a single document. We learn an LDA model by finding the parameters α and β such that the log marginal likelihood of the entire database is maximized. Since Eqn. 1 cannot be solved directly, model parameters are estimated by variational inference [4].
Given the learned corpus parameters α and β, the LDA model allows us to assign probabilities to data outside the training corpus by maximizing the log marginal likelihood of the respective document. Thus we may learn the LDA corpus level parameters on a subset of the image database (in order to reduce total training time) and then assign probability distributions to all images. This is one of the advantages of the fully generative LDA topic model compared to the pLSA aspect model. In the pLSA model there exists no direct way to assign probabilities to unseen documents. Additionaly the LDA overcomes some overfitting problems of the pLSA which occur due to its large set of parameters that are directly linked to the training set [4].
Several extensions of the LDA model have been proposed [3, 14].

2.2 Image Similarity Measures

Once we train an LDA model and we compute a probabilistic representation for each image in the database, we need to define an image similarity measures in order to perform image retrieval. In this work, we focus on the task of query-by-example, thus searching in the database for the most similar items to a given query image. The topic mixture θ for each image indicates to what degree a certain topic is contained in the respective image. Based on the topic mixtures, we look at four different ways to measure similarity and evaluate these measures experimentally in Section 3.
First the similarity between two images I_a and I_b can be measured by calculating the cosine similarity between the topic distributions $\mathbf{P}(\mathbf{z}|\theta^{\mathbf{a}}, \alpha)$ and $\mathbf{P}(\mathbf{z}|\theta^{\mathbf{b}}, \alpha)$. The cosine $\cos(\mathbf{a}, \mathbf{b})$ between two vectors $\mathbf{a}$ and $\mathbf{b}$ is popluar in text retrieval [1] and is defined by:

$$cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|} \quad (2)$$

A second possibility to measure image similarity is the use of the $L1$ distance between two topic distributions. The $L1$ distance between to K dimensional vectors $\mathbf{a}$ and $\mathbf{b}$ is given by:

$$L1(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{K} |a_i - b_i| \quad (3)$$

The third similarity measure that we study is the symmetrized Jensen-Shannon divergence $JS(\mathbf{P}(\mathbf{z}|\theta^{\mathbf{a}}, \alpha), \mathbf{P}(\mathbf{z}|\theta^{\mathbf{b}}, \alpha))$ between the topic distributions of two images. The JS measure is based on the discrete Kullback Leibler divergence $KL(\mathbf{P}(\mathbf{z}|\theta^{\mathbf{a}}, \alpha), \mathbf{P}(\mathbf{z}|\theta^{\mathbf{b}}, \alpha)$:

$$JS(\mathbf{a}, \mathbf{b}) = \frac{1}{2}(KL(\mathbf{a}, \frac{\mathbf{a}+\mathbf{b}}{2}) + KL(\mathbf{b}, \frac{\mathbf{a}+\mathbf{b}}{2})) \quad (4)$$

where

$$KL(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{K} a_i \log \frac{a_i}{b_i} \quad (5)$$

The fourth measure is adopted from language based information retrieval. Here, each document is indexed by the likelihood of its model generating the query document, i.e. the most relevant documents are the ones whose model maximizes the conditional probability on the query terms. In content-based image retrieval, a query image can be pesented as a sequence of visual words $\mathbf{w_a}$ and the above mentioned likelihood can be written as:

$$P(\mathbf{w_a}|M_b) = \prod_{i=1}^{N_d} P(w_i^a|M_b) \quad (6)$$

where M_b is the model of an image I_b and N_d the total number of detected visual words in image I_a.
Wei and Croft [17] combine the LDA model and the unigram model with Dirichlet smoothing to estimate the terms $P(w_i^a|M_b)$:

$$P(w_i^a|M_b) = \lambda \cdot P_u(w_i^a|M_b^u) + (1-\lambda) \cdot P_{LDA}(w_i^a|M_b^{lda}) \quad (7)$$

Category	OR list of tags	# of images
1	wildlife animal animals cat cats	28509
2	dog dogs	24660
3	bird birds	20908
4	flower flowers	25457
5	graffiti	21888
6	sign signs	14333
7	surf surfing	29552
8	night	33142
9	food	18602
10	building buildings	16826
11	goldengate goldengatebridge	23803
12	baseball	12372
	Total # of images (Note images may have multiple tags)	246,348

Table 1: Image database and its categories used for experiments

where $P_u(w_i^a|M_b^u)$ is specified by the unigram document model with Dirichlet smoothing according to [18]:

$$P_u(w_i^a|M_b^u) = \frac{N_d}{N_d+\mu}P_{ML}(w_i^a|M_b^u) + (1-\frac{N_d}{N_d+\mu})P_{ML}(w_i^a|D) \quad (8)$$

Here D denotes the entire set of images in the database and μ the Dirichlet prior. The term $P_{LDA}(w_i^a|M_b^{lda})$ in Eqn. 7 refers to the probability of a visual word w_i^a in image I_a given the LDA topic model M_b^{lda} of image I_b:

$$P_{LDA}(w_i^a|M_b^{lda}) = P_{LDA}(w_i^a|\alpha,\theta^b,\beta) = \sum_{j=1}^{K} P(w_i^a|z_j,\beta)\cdot P(z_j|\theta^b,\alpha) \quad (9)$$

3. EXPERIMENTAL RESULTS

The objective of example-based image retrieval is to obtain images with content similar to the given sample image. We evaluate retrieval results based on the judgments of several test users about the visual similarity of the retrieved images with respect to the query image.

All experiments are performed on a database consisting of approximately 246,000 images. The images were selected from all public Flickr images uploaded prior to Sep. 2006 and labeled as *geotagged* together with one of the following tags: *sanfancisco, beach,* and *tokyo*. Of these images only images having at least one of the following tags were kept: *wildlife, animal, animals, cat, cats, dog, dogs, bird, birds, flower, flowers, graffiti, sign, signs, surf, surfing, night, food, building, buildings, goldengate, goldengatebridge, baseball*. The images can thus be grouped into 12 categories as shown in Table 1.

The preselection of a subset of images from the entire Flickr database based on tags is needed as Flickr is a repository with hundreds of millions of images. However, it should be noted, that indexing purely based on tags is not sufficient as the tags are a very noisy indication of the content shown in the images.

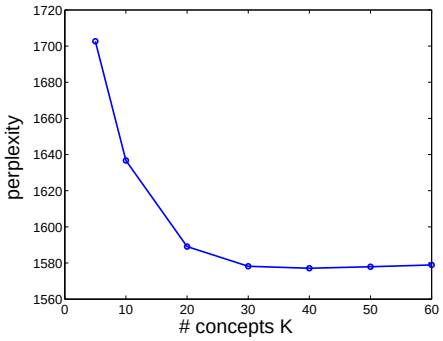


Figure 2: Perplexity vs. number of topics K

We computed the visual vocabulary from 12 randomly selected non-overlapping subsets each consisting of 500,000 local features. Each of those subsets produces 200 visual words giving a total vocabulary size of 2400 visual words.

3.1 Parameter Settings

The first step in evaluating the retrieval system is to determine suitable parameters for the LDA model, such as the number of training images as well as the number of topics K. Thus a suitable measure to assess the performance with respect to different parameter settings is needed. The *perplexity* is frequently used to assess the performance of language models and to evaluate LDA models in the context of document modeling [4]. It measures the performance of the model on a held out dataset D_{test} and is defined by:

$$per(D_{test}) = exp\left\{-\frac{\sum_{d=1}^{M}\log p(\mathbf{w_d})}{\sum_{d=1}^{M}N_d}\right\} \quad (10)$$

This measure decreases monotonically in the likelihood of the test data, thus lower values indicate better modeling performance.

In order to evaluate the influence of the choice of the number of hidden topics, we trained an LDA model on a subset of 50,000 images using different numbers of aspects. The perplexity is then calculated on a previously unseen test set of 25,000 images. Figure 2 shows the perplexity plotted against the number of hidden aspects K. One can see that the perplexity decreases with an increasing number of topics. If the number of topics is small, i.e. $K < 30$, the perplexity grows rapidly indicating that the model does not fit the unseen training data. For $K \geq 30$ the perplexity is almost constant. We need a rich image description for our retrieval task, thus we will set $K = 50$ in our experiments.

Figure 3 shows the perplexity for different sizes of the training set, i.e. the number of images in the training set is varied. The number of topics is fixed to $K = 50$ in order to evaluate the change of the perplexity with respect to the number of images used for training the LDA corpus level parameters. Perplexity is again calculated for each setting based on a perviously unseen test set consisting of 25,000 images. The perplexity decreases with an increasing number of training samples and is approximately constant for training set sizes above $20,000$ images. However, the decrease in perplexity is not as fast as the decrease based on the choice of the number of topics, thus no definite conclusion about the appropriate size of training samples can be drawn. The

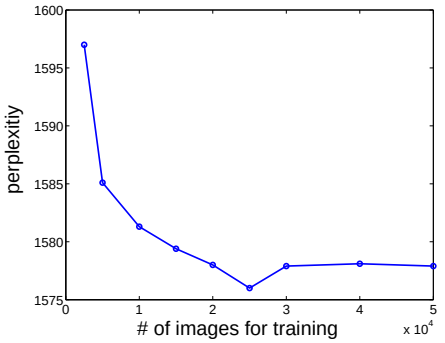


Figure 3: Perplexity vs. number of training samples

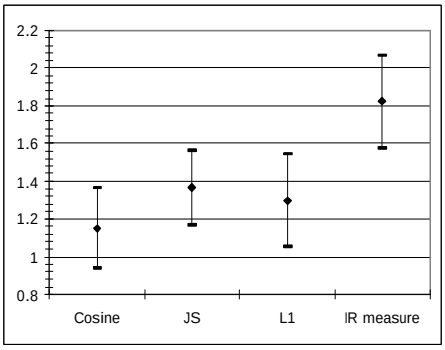


Figure 4: User preferences for the four image similarity measures using the LDA image representation

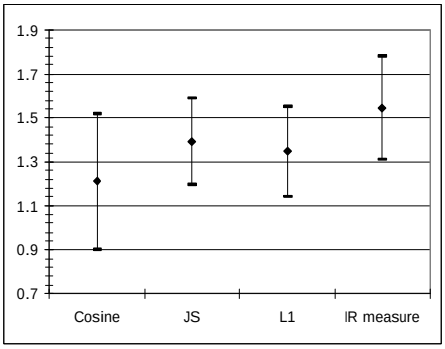


Figure 5: User preferences for the four image similarity measures using the pLSA image representation

appropriate number of images used to train the LDA model may also depend on other parameters such as the choice of the maximum number of iteration in the variational inference part as well as the number of topics and the size of the vocabulary, respectively. It is still important to notice that in our tests it does not seem to be necessary to perform LDA model computation on the entire database, which is a huge advantage in large-scale databases. It also enables adding novel images to the database without relearning the LDA corpus level parameters as long as they show already learned topics.

3.2 Different Similarity Measures

We decribed different similarity measures for the LDA-based image representation in Section 2.2. Here we evaluate their effects on the image retrieval task, with the number of topics in the LDA model set to 50 and the model trained on 50,000 images. Once we compute the model, we assign probabilities to all images by maximizing the log marginal likelihood of the respective document (see Section 2.1). The parameters μ and λ of the information retrieval based distance measure are set to 50 and 0.2, respectively.

We judge the effect of the similarity measures on the retrieval results by users: We selected five query images per category at random resulting in a total of 60 query images for the experiments. For each query image the 19 most similar images derived by the four different measures are presented to the users. The test users were asked to judge the retrieval results by putting them in an order from best to worst by assigning 3 points to the best technique, 2 points to the second best, and 1 and 0 to the second worst and worst performing technique, respectively. We compute the average score for each method over all 60 images. It should be noted that sometimes the performance of all four techniques were not satisfying at all. In those cases the user could assign 0 points to all similarity measures. We allow a corresponding procedure in cases where all four techniques produced perfect results: all techniques could earn 3 points.

We depict the resulting mean scores over 10 test users in Figure 4. The vertical bars mark the standard deviation of the test users' scores. The best performing distance measure is the probability measure adopted from information retrieval (Eqn. 6) [17]. This indicates that retrieval based on the topic distribution is enhanced by also taking word distributions into account. Note, that the word probability calculated based on the unigram model is assigned only a small weight of 0.2 whereas the word probability based on the LDA model is assigned large weight (0.8).

Out of the three similarty measures based on only the topic distributions, the Jensen-Shannon divergence performs best, followed by the L1 distance. If image retrieval is performed on large-scale databases the probability measure from information retrieval may be too time consuming and dimensionality reduction in image representation is important. In this case one should also consider the second best approach, the Jensen-Shannon divergence. As word occurences are soley needed to build the LDA representation, only the low dimensional topic distribution needs to be stored and processed for the retrieval task. This allows us to search even larger databases in reasonable time.

3.3 pLSA versus LDA

In our earlier study [9] we used a latent aspect model to represent images in the context of a retrieval-by-example task. The work combined the topic vector produced by the pLSA model for each image with the cosine distance measure and found that this approach outperformed the pure visual word-occurrence vectors as well as color coherence vectors [11]. In this work we use the same real world database for evaluation purposes.

In order to determine the most appropriate image representation for the studied retrieval task, the results obtained using LDA-based image features should be compared to those derived using the pLSA-based image representation. Since the previous section shows that retrieval performance depends on the distance measure used, the most suitable dis-

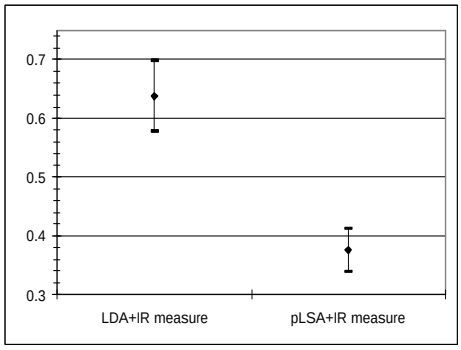


Figure 6: User preferences for the comparsion between the retrieval approach using LDA image features and the approach using pLSA image features

tance measure for the pLSA representation needs to be identified first. Thus, all four similarity measures described in Section 2.2 are applied to the pLSA-based image representation.[1] Results on 60 query images are then judged by test users as described in the previous section. Mean scores and standard deviations of 10 test users are shown in Figure 5. A similar result as for the LDA image features is observed for the pLSA-based image representation (see also Section 3.2). The IR measure outperforms all other similarity measures, followed by the Jensen-Shannon divergence. The cosine distance shows the worst performance. It can also be seen that the resulting scores are more consistent, thus not showing as large differences between the distance measures as we obtained using the LDA image features.

As we obtain relative scores only for the comparsion of similarity measures, we still do not know which image features are more appropriate for the retrieval task. Therefore, we compare the retrieval performance of LDA and pLSA-based features using the best performing similarity measure – the IR measure (Eqn. 6) – by using 60 images evaluated by 10 users. Since we compare only two techniques in the experiment, test users judge the retrieval results of each query image by assigning 1 point to the better performing method and 0 points to the other method. Mean scores and standard deviations are depicted in Figure 6. It can be clearly seen that the score of the pLSA-based image representation is significantly lower than the results for the LDA-based image representation. Thus, we conclude that the LDA-based image representation studied in this work is more suited for the image retrieval task on a large real world database than the pLSA-based image features.

3.4 Results

Finally we show some retrieval results obtained by the proposed LDA-based system in Figure 7. As one can see, in the top seven rows the systems performs very well. The following rows show queries were the returned results are suboptimal, especially in the last row the systems fails completely. Displayed results are obtained using different similarity measures.

[1]The number of topics in the pLSA model is set to 48.

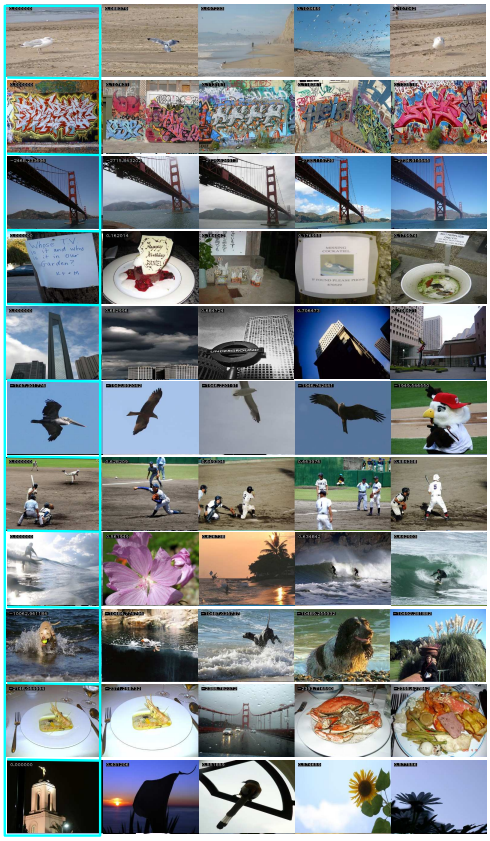


Figure 7: Retrieval results obtained by our LDA-based system. The left most image in each row shows the query image, the four images to the right show the most relevant images retrieved.

4. ACTIVE LEARNING

So far retrieval results are obtained in a completely unsupervised manner. In this section, active learning is deployed to improve retrieval results. In active learning the user interactively informs the system about the concept he/she is looking for. The system poses 'questions' to the user, which the user must answer in order to provide feedback to the system about his/her actual search goal. Questioning is performed by asking the user to label an image or a set of images as *relevant* or *irrelevant*. As users expect the system to capture their desired concept effectively, i.e. quickly and accurately, the main issue in designing such systems is finding the most informative instances to present for labeling purposes to the user. In this work we combine the support vector machine (SVM) based active learning approach [15] on the LDA-based image representation with a simple preprocessing scheme to effectively prune the image candidate space.

4.1 SVM-based Active Learning

Tong and Chang [15] proposed active learning with support vector machines (SVM) by regarding the task of learning a target concept as the task of separating the relevant images from the irrelevant ones by learning an SVM binary classifier, i.e. a hyper plane in some high dimensional space. The presented active learning method works as follows: An SVM

classifier is trained in each query round. In the first query round the algorithm is initialized with one relevant and one irrelevant image and the user labels a randomly selected set of T images. In each following round the T most informative images are presented to the user for labeling. The most informative images are defined as the closest images to the current hyper plane according to the so called 'simple method'. After a number of relevance feedback rounds, the most relevant images are presented to the user as the query result. The binary SVM classifier subdivides the space by the hyper plane in two sets, relevant and irrelevant images and thus the most relevant images are those that are farthest from the current SVM boundary in the kernel space and on the right side of the hyper plane.

In order to apply this algorithm to images, each image needs to be presented as a vector. We propose to represent the images in the database by their $\mathbf{P}(\mathbf{z}|\theta, \alpha)$ distributions, thus combining LDA image representation and SVM active learning.

The active learning algorithm works well for small databases with carefully selected images. Problems arise when applying this algorithm to large-scale databases. First, the user needs to find at least one positive query image to initialize the algorithm. Fortunately in this work the query by-example task is considered and thus the example image can be used to initialize the algorithm. A second problem arises due to the number of images showing the desired content with respect to the total number of images in the database. If this fraction is very small (as it usually is in large-scale databases), active search is aggravated.

In order to solve this problem, a preprocessing step is performed before starting the active learning algorithm. This preprocessing step aims to reduce the total amount of images in the database while at the same time keeping images that likely contain the desired concept, i.e. the active learning algorithm will not work on the entire database of 246,000 images but only on a preselected subset of images. As a convenient side effect of preprocessing, computation time of each query round is reduced as the algorithm is running on a smaller dataset making active search faster.

The proposed data selection approach takes advantage of the learned LDA image representation: We choose a subset of R images for active learning based on the prior detected relevance to the query image. Relevance is defined by similarity based on the LDA image representation and the distance measures discussed in Section 2.2. This makes intuitive sense as an LDA-based image representation models the image content by topic assignment and thus images having completely different topic distributions are unlikely to match the desired user's concept.

4.2 Experimental Results

For the evaluation of the active learning algorithm based on LDA image features the parameters in the experiment are set as follows: images are represented by their topic distribution, which is learned from a 50 topics LDA model. We used a radial basis function (RBF) kernel with $\alpha = 0.01$ in the SVM and we set the number of query images T presented to the user in each query round to 20. We chose the parameter R, the size of the preselected subset, to be 20,000. This ensures a sufficient downsizing from the original total amount of images while at the same time keeping an adequate number of images likely containing the desired content. The subset of R images is determined by applying the $L1$ distance on the topic distributions.

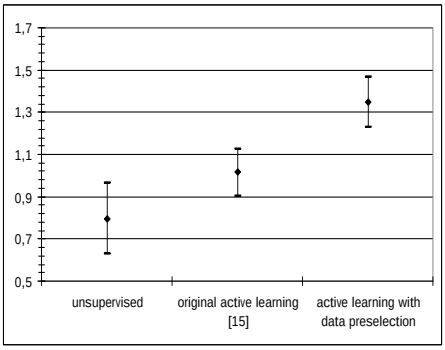


Figure 8: User preferences for the comparsion between the two active learning approaches and the unsupervised approach

The results of the active learning algorithm with pre-filtering are compared to the results obtained by the active learning algorithm without pre-filtering [15] and the results from unsupervised retrieval using the IR similarity measure. Evaluation is again performed through user studies. 25 sample query images are chosen from the pool of 60 images used for the evaluation in Section 3. As a common user will most likely perform no more than three to four query rounds we presented the 19 most relevant images to the given query concept after three rounds of active learning to the test users. Test users compare the results of all three methods and the best performing method earns 2 points, whereas the second best and worst receive 1 and 0 points, respectively. The mean over all 25 images is then calculated and the results over all 10 test users are depicted in Figure 8. The results show that active learning clearly improves the results compared to the complete unsupervised retrieval. Moreover, an additional improvement over the original active learning algorithm [15] can be achieved by using pre-filtering (i.e., data preselection).

In Figure 9 some sample results showing the effectiveness of the presented active learning approach are depicted. Three pairs of 20 images are displayed, each pair showing the query image and the nine most relevant images found using the unsupervised algorithm evaluated in Section 3 (top) and after three rounds of active learning with data preselection (bottom). Green dots mark images showing the correct content, red dot mark incorrectly retrieved images. Clearly an improvement of the results by active learning can be noticed.

5. CONCLUSIONS

This work studies the representation of images by Latent Dirichlet Allocation (LDA) models in the context of query-by-example retrieval on a large real world image database consisting of more than 246,000 images. Results show that the approach performs well. The combination of LDA-based image representation with an appropriate similarity measure outperforms previous approaches such as a pLSA-based image representation. We found that a similarity measure developed for information retrieval and based on probabilities gives the best retrieval results. We examined the application of an active learning algorithm on the LDA-based image features and proposed a novel data subset selection scheme for

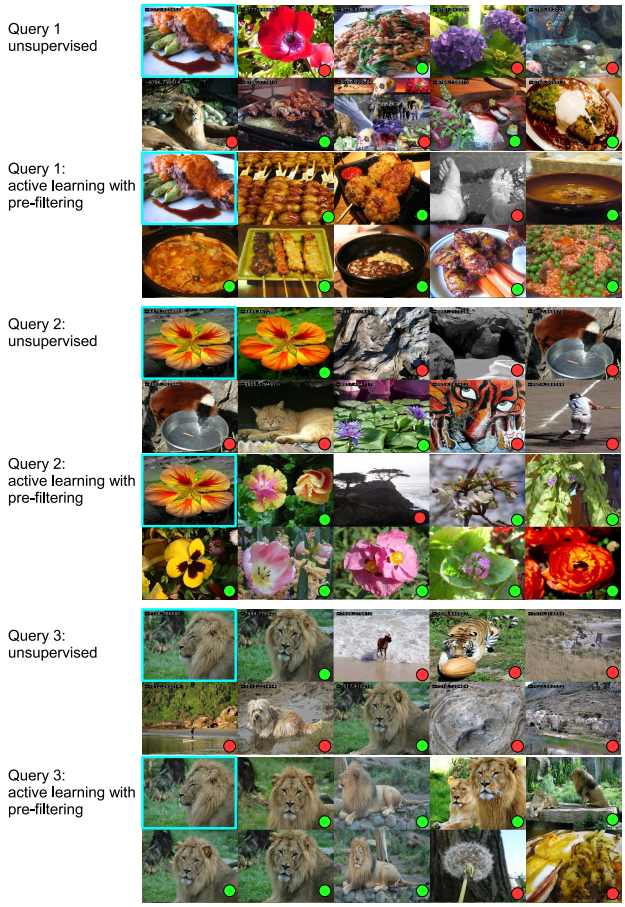


Figure 9: Retrieval results: Each image pair shows the results obtained by the unsupervised algorithm (top) and by active learning with pre-filtering (bottom)

retrieval in large databases. Future work will verify the results using a larger number of users and we will incorporate different types of image features.

6. REFERENCES

[1] R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[2] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, 2003.

[3] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, New York, NY, USA, 2003. ACM Press.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[5] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *Proceedings of the European Conference on Computer Vision*, 2006.

[6] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 1816–1823, Washington, DC, USA, 2005. IEEE Computer Society.

[7] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1-2):177–196, 2001.

[8] F.-F. Li and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 524–531, Washington, DC, USA, 2005. IEEE Computer Society.

[9] R. Lienhart and M. Slaney. pLSA on large scale image databases. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[11] G. Pass, R. Zabih, and J. Miller. Comparing images using color coherence vectors. In *MULTIMEDIA '96: Proceedings of the fourth ACM international conference on Multimedia*, pages 65–73, New York, NY, USA, 1996. ACM Press.

[12] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 883–890, Washington, DC, USA, 2005. IEEE Computer Society.

[13] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *International Conference on Computer Vision (ICCV 2005)*, 2005.

[14] E. B. Sudderth, A. B. Torralba, W. T. Freeman, and A. S. Willsky. Describing visual scenes using transformed dirichlet processes. In *NIPS*, 2005.

[15] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *MULTIMEDIA '01: Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118, New York, NY, USA, 2001. ACM Press.

[16] G. Wang, Y. Zhang, and L. Fei-Fei. Using dependent regions for object categorization in a generative framework. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, Washington, DC, USA, 2006. IEEE Computer Society.

[17] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, New York, NY, USA, 2006. ACM Press.

[18] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342, New York, NY, USA, 2001. ACM Press.